REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
02-09-2017	Final Report	30-Sep-2010 - 29-Sep-2016

4. TITLE AND SUBTITLE
Final Report: The Linguistic-Core Approach to Structured Translation and Analysis of Low-Resource Languages

5a. CONTRACT NUMBER
W911NF-10-1-0533

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Carnegie Mellon University
5000 Forbes Avenue

Pittsburgh, PA 15213 -3589

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58138-MA-MUR.77

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UU	UU	UU	UU		Jaime Carbonell
					19b. TELEPHONE NUMBER
					412-268-7279

RPPR Final Report
as of 05-Sep-2017

Agency Code:

Proposal Number: 58138MAMUR

Agreement Number: W911NF-10-1-0533

INVESTIGATOR(S):

Name: Jaime Carbonell
Email: jgc@cs.cmu.edu
Phone Number: 4122687279
Principal: Y

Name: Lori Levin Ph.D.
Email: lsl@cs.cmu.edu
Phone Number: 4122687130
Principal: N

Name: Noah Smith
Email: nasmith@cs.cmu.edu
Phone Number: 4122684963
Principal: N

Name: Stephan Vogel
Email: stephan.vogel@cs.cmu.edu
Phone Number: 4122684526
Principal: N

Organization: **Carnegie Mellon University**
Address: 5000 Forbes Avenue, Pittsburgh, PA 152133589
Country: USA

DUNS Number: 052184116 EIN: 250969449
Report Date: 29-Dec-2016 Date Received: 02-Sep-2017
Final Report for Period Beginning 30-Sep-2010 and Ending 29-Sep-2016
Title: The Linguistic-Core Approach to Structured Translation and Analysis of Low-Resource Languages
Begin Performance Period: 30-Sep-2010 **End Performance Period:** 29-Sep-2016
Report Term: 0-Other
Submitted By: Lori Levin Email: lsl@cs.cmu.edu
Phone: (412) 268-7130
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: **STEM Participants:**

Major Goals: The project goals were to conduct research on machine translation and textual analysis for low-resource languages, focusing (but not exclusively) on African languages such as Kinyarwanda, Malagasy, and Swahili.

Accomplishments: Accomplishments are described in the uploaded report. The MURI group collected resources in three African languages (Swahili, Kinyarwanda, and Malagasy) and built machine translation and textual analysis systems for these languages. The group also conducted research on techniques for MT and TA of low resource languages such as multi-lingual transfer and compact representations. Significant progress was made on graph transduction techniques, monolingual decipherment, modeling lexical borrowing, and semantics based MT and TA.

Training Opportunities: Nothing to Report

Results Dissemination: We have published our results at many of the top conferences in our field.

Honors and Awards: Jaime Carbonell, Okawa Prize, 2015

Kevin Knight, Fellow, Association for Computational Linguistics, 2014

RPPR Final Report
as of 05-Sep-2017

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Faculty
Participant: Jaime Carbonell
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Noah Smith
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Lorraine Levin
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Regina Barzilay
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Kevin Knight
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: David Chiang

Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Vamsi Krishna Ambati
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jeffrey Flanigan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Sanjika Hewavitharana
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Daniel Mills
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yevgeni Berzak
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Satchuthananthavale Branavan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Georgiana Gane
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Nathaniel Kushman
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yoong Keok Lee
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Christina Sauper
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Kyle Jerro
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Vijay John
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Evelyn Richter
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Christopher Dyer
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Victoria Fossum
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Vamsi Krishna Ambati
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Waleed Ammar
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

RPPR Final Report
as of 05-Sep-2017

Participant Type: Graduate Student (research assistant)
Participant: Victor Aurelien Chahuneau
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jeffrey Flanigan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Kevin Gimpel
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Daniel Mills
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Aaron B Phillips
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Christopher Dyer
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Jaime Carbonell
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Noah Smith
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Lorraine Levin
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: David Chiang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Kevin Knight
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Jason Baldridge
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:

Other Collaborators:

Participant Type: Faculty
Participant: Regina Barzilay
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Cai Shu
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Licheng Fang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Vijay John
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Kyle Jerro
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yevgeni Berzak
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:

National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Satchuthanathavale Branavan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Georgiana Gane
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Nathaniel Kushman
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yoong Keok Lee
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Tahira Naseem
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Christina Sauper
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:

International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yuan Zhang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Victoria Fossum
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Carbonell, Jaime
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Smith, Noah
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Levin, Lorraine
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Dyer, Christopher
Person Months Worked: **Funding Support:**
Project Contribution:

International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Barzilay, Regina
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Knight, Kevin
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Chiang, David
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Baldridge, Jason
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Ammar, Waleed
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Chahuneau, Victor Aurelien Jacques
Person Months Worked: **Funding Support:**

Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Flanigan, Jeffrey
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Tsvetkov, Yulia
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Berzak, Yevgeni
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Hynes, Zachary
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Kushman, Nathaniel
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Lee, Yoong Keok

Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Lei, Tao
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Naseem, Tahira
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Zhang, Yuan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Shu, Cai
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Qing, Dou
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)

Participant: Lee, Chia-Ying
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Mielens, Jason
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Sun, Liang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Bhatia, Archna
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Carbonell, Jaime
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Smith, Noah
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Levin, Lorraine
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Dyer, Christopher
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Barzilay, Regina
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Knight, Kevin
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Chiang, David
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Ammar, Waleed
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Flanigan, Jeffrey
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Tsvetkov, Yulia
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Berzak, Yevgeni
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Hynes, Zachary
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Lee, Yoong Keok
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Deri, Aliya
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Gao, Yang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Pourdamghani, Nima
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Qing, Dou
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Lee, Chia-Ying
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Schamper, Julian
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Vaswani, Ashish
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:

Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Bhatia, Archna
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Feng, Yang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Baldridge, Jason
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Mielens, Jason
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Sun, Liang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Garrette, Dan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:

National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Jaime Carbonell
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Noah Smith
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Chris Dyer
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Lori Levin
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Kevin Knight
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: David Chiang
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:

International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Regina Barzilay
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Faculty
Participant: Jason Baldridge
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Waleed Ammar
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jeffrey Flanigan
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yulia Tsvetkov
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Aliya Deri
Person Months Worked: **Funding Support:**
Project Contribution:

International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yang Gao
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Nima Pourdamghani
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Dou Qing
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Chia-Ying Lee
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Julian Schamper
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Ashish Vaswani
Person Months Worked: **Funding Support:**

Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yevgeni Berzak
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Zachary Hynes
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yoong Keok Lee
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jason Mielens
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Liang Sun
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Dan Garrette

Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Archna Bhatia
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Yang Feng
Person Months Worked: **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member:
Other Collaborators:

The Linguistic-Core Approach to Structured Translation and Analysis of Low-Resource Languages

Final Report 2011-2016

Jaime Carbonell, PI
Jason Baldridge, Regina Barzilay, David Chiang, Chris Dyer,
Kevin Knight, Lori Levin, and Noah A. Smith

Abstract

The Linguistic Core MURI project focused on machine translation (MT) and textual analysis (TA) engines for low resource languages. We produced systems that can be trained with less data by using knowledge-rich linguistic priors, linguistic corpus annotation, monolingual corpora, techniques for cross-lingual training of NLP systems, and compact representations that allow for generalization over small amounts of data. Our research activities ranged from data collection and annotation to the design, development, and evaluation of algorithms and models for text analysis and machine translation. Our work addressed three focus languages from Africa (Kinyarwanda, Malagasy, and Swahili), but we also piloted many techniques on a variety of other languages. This report covers work that was done in the five years of the project and the sixth year extension.

Contents

1 Overview

This report highlights six years of research on MT and textual analysis (TA) for low-resource languages. Some of our work focused on three specific African languages (Kinyarwanda, Malagasy, and Swahili). Other work focused on low-resource scenarios in other languages, and some focused on new, high-impact techniques.

2 Target languages: Swahili, Kinyarwanda, and Malagasy

The focus languages for this project were Kinyarwanda, Malagasy, Swahili, and Yoruba. We report here on work that we did on Kinyarwanda, Malagasy, and Swahili. We did not pursue Yoruba because preliminary efforts showed that it was much less feasible than the others, although some Yoruba data is included in our repository of data and tools.

2.1 Data Resources in the Target Languages

We collected and processed significant text data for the target languages. For each language, we selected and cleaned the texts, tokenized the texts, and aligned parallel text where possible. Addition processing and annotation is described in §2.2 and §2.3.

2.1.1 Malagasy

- Malagasy parallel text, collected from `http://www.lakroa.mg/` and `http://www.lagazette-dgi.com/` and interlinear glossed texts, and translated by volunteer informants `https://github.com/ldmt-muri/muri-data/tree/master/mlg/orig`
- Malagasy parallel text from the GlobalVoices project, collected and aligned `http://www.cs.cmu.edu/~ark/global-voices/`
- Malagasy parallel text from an aligned Bible translation `https://github.com/ldmt-muri/muri-data/tree/master/mlg/orig/mlg_bible`

2.1.2 Kinyarwanda

- Monolingual Kinyarwanda text from the Izuba newspaper `https://github.com/ldmt-muri/muri-data/tree/master/kin/orig/izuba`

- Kinyarwanda parallel text from the Kigali Genocide Memorial Center (KGMC), translated by KGMC translators `https://github.com/ldmt-muri/muri-data/tree/master/kin/orig/kgmc`

- Kinyarwanda parallel text from the BBC, translated by volunteer informants `https://github.com/ldmt-muri/muri-data/tree/master/kin/orig/bbc`

2.1.3 Swahili

- Monolingual Swahili text from newswire `https://github.com/ldmt-muri/muri-data/tree/master/swa2/data/news` and Wikipedia `https://github.com/ldmt-muri/muri-data/tree/master/swa2/data/wiki`.

- Parallel Swahili text from the GlobalVoices project `https://github.com/ldmt-muri/muri-data/tree/master/swa2/data/globalvoices_raw`

2.1.4 Yoruba

- Monolingual Yoruba text from Wikipedia `https://github.com/ldmt-muri/muri-data/tree/master/yor/wiki`

2.1.5 Multilingual

- The URIEL database: a collection of typological, geographical, and phylogenetic features (collected from WALS [21], SSWL [15], PHOIBLE [50], and [40], text-mined from Ethnologue [46], and predicted by kNN regression) for Yoruba, Swahili, Kinyarwanda, and Malagasy (among other languages) [48] `http://www.cs.cmu.edu/~dmortens/uriel.html`

2.2 Human-Annotated in the Target Languages

We prepared human-annotated subsets of the above data for supervised learning, and for evaluation of rule-based systems, concentrating on Kinyarwanda and Malagasy.

2.2.1 Part-of-speech annotation

- POS annotation of the Malagasy Global Voices parallel text https://github.com/ldmt-muri/muri-data/tree/master/mlg/https://github.com/ldmt-muri/muri-data/tree/master/mlg/tagged
- POS annotation of the Kinyarwanda KGMC parallel text https://github.com/ldmt-muri/muri-data/tree/master/kin/tagged

2.2.2 Morphological annotation

- Morphological analysis of Malagasy news text https://github.com/ldmt-muri/muri-data/tree/master/mlg/morph
- Morphological analysis of Kinyarwanda KGMC and news text https://github.com/ldmt-muri/muri-data/tree/master/kin/morph

2.2.3 Graph Fragment Language annotation

- GFL annotation of the Malagasy Global Voices parallel text https://github.com/ldmt-muri/muri-data/tree/master/mlg/gfl/global_voices
- GLF annotation of the Kinyarwanda KGMC parallel text https://github.com/ldmt-muri/muri-data/tree/master/kin/gfl/kgmc

2.3 Textual Analysis of Target Languages

We made a number of textual analyzers—e.g., morphological parsers, part-of-speech taggers, and dependency parsers—using supervised, unsupervised, and rule-based techniques.

2.3.1 Morphological analysis

- MorphoChain, unsupervised morphological analyzer for Swahili (among other languages) https://github.com/karthikncode/MorphoChain/
- LLABSWA, a FST morphological transducer for Swahili https://pwlittell.com/resources/llabswa.zip
- kin-morph-fst, a FST morphological transducer for Kinyarwanda https://github.com/ldmt-muri/kin-morph-fst

- Malagasy morphological analysis: we acquired an existing FST morphological transducer for Malagasy and adapted it for use in semi-automatic annotation of morphological structure `https://github.com/ldmt-muri/muri-data/tree/master/mlg/morph/FST`

2.3.2 Dependency parsing

- RBGParser, Multi-lingual dependency parser for Kinyarwanda and Malagasy (among other languages) `https://github.com/taolei87/RBGParser/`

- Supervised dependency parsers and POS taggers for Kinyarwanda and Malagasy `http://www.ark.cs.cmu.edu/TurboParser/`

- Gibbs parser for Kinyarwanda and Malagasy `https://github.com/jmielens/gibbs-pcfg-2014`

2.3.3 Part-of-speech tagging

- POS tagger for Kinyarwanda and Malagasy ([32]; [34]) `https://github.com/dhgarrette/low-resource-pos-tagging-2014`

- CRF Autoencoder framework for POS tagging in Kinyarwanda and Malagasy (among other tasks and languages) `https://github.com/ldmt-muri/alignment-with-openfst/tree/master`

2.3.4 Lexicon induction

- MonoGiza: a program that produces a word-to-word probabilistic dictionary ("t-table") from non-parallel text, for Malagasy (among other languages) `http://www.isi.edu/natural-language/software/`

2.4 MT in Target Languages

2.4.1 Year 1

Using the collected data for Kinyarwanda and Malagasy, we were able to produce preliminary MT systems. CMU produced Kinyarwanda and Malagasy systems and implemented improvements using CRF word alignments. USC/ISI built end-to-end translation systems for Kinyarwanda-English, Malagasy-English, English-Kinyarwanda, and English-Malagasy. For each direction, they built a Hiero MT system, and for each English-target direction, they built a syntax-based MT system. CMU and USC compared results and collaborated on improvements.

Data	Bleu (↑)	TER (↓)	OOV Rate (↓)
Year 1 data (19th c. Bible only)	8.4	82.8	9.8
Year 2 data (Global Voices through June 2011)	18.3	67.6	2.2
Year 2 data (Global Voices through June 2012)	19.1	66.9	1.5

Table 1: Malagasy-English translation results showing the benefit of more data . OOV rate is per-sentence

reference	Students in Bogota participating in a protest.
year 1	Disciples in Bogota takes part in the opposition.
year 2	Students in Bogota participating in the protest.
reference	Am I wrong in questioning the credibility of #wikileaks?
year 1	I lack ve new if they ask touches the righteousness of the #wikileaks?
year 2	Am I wrong if wonder about the stability of the new #wikileaks?
reference	Morocco: war on press continues
year 1	Maraoka: mitohy battle in the gazety
year 2	Morocco: the fight against the journalists continues

Table 2: Malagasy-English translation output for models trained using only the year 1 data and the full year 2 dataset (first and last lines in Table 1. Note the improvements in word choice in the output; where Biblical words like *disciples* and *righteousness* and Malagasy words were included in year 1 output, they have been replaced by more suitable vocabulary. Of course, further room for improvement is apparent.

2.4.2 Year 2

We implemented and evaluated a range of systems involving translation into and out of the focus languages. Key results are shown in Tables 1-3. Table 1 shows the huge benefit of additional data on translation performance (automatic evaluation); examples comparing output from the year 1 system and the full-data system are given in Table 2. The benefit of quasi-synchronous phrase dependency model (developed during year 1; [36]) is demonstrated for English-Malagasy translation in Table 1.

2.4.3 Year 3

[28] developed a discriminative learning algorithm for machine translation systems that targets mixes of frequently and infrequently occurring features, estimating the sparse features on large amounts of parallel training data. Unlike previous ap-

Translation Model	Bleu (↑)
Phrase-based (Koehn et al., 2003)	15.1
Hierarchical phrase-based (Chiang, 2005)	15.0
Quasi-synchronous phrase dependencies (Gimpel and Smith, 2011)	15.5

Table 3: English-Malagasy translation results.

System	Dev	Test
MERT Baseline	19:80:3	17:70:2
Our algorithm	20:50:1	18:40:2

Table 4: MalagasyEnglish translation quality with new discriminative training algorithm and sparse, rule-indicator features.

proaches that support learning extremely sparse features, high quality development data continues (which enables better estimation of translation quality) continues to be used to determine the relative contributions of the sparse features in the model along side less-sparse features.

2.4.4 Year 4

Decipherment for Malagasy translation. Inspired by previous work where decipherment is used to improve machine translation, we proposed a new idea to combine word alignment and decipherment into a single learning process ([19]). We used EM to estimate the model parameters, not only to maximize the probability of parallel corpus, but also the monolingual corpus. We applied our approach to improve Malagasy-English machine translation, where only a small amount of parallel data is available.

In Table 5, we show translation accuracy improvements between 0.9 and 2.1 BLEU points over a strong baseline. "Pipelined Decipherment" uses a small parallel text to seed decipherment of monolingual Malagasy, which results in new bilingual dictionary entries. These are fed back into an MT system trained on parallel text. "Joint Decipherment" is our new method, which processes parallel and monolingual data simultaneously, so that each informs the analysis of the other, resulting in cleaner bilingual dictionaries and better translation.

2.4.5 Year 5

Malagasy: We published the last in a series of papers on Malagasy decipherment, i.e., how to learn a word-to-word mapping between Malagasy and English without

System	Tune BLEU GlobalVoices	Test BLEU GlobalVoices	Test BLEU Web text
Phrase-Based MT	18.5	17.1	7.7
Pipelined Decipherment	18.5	17.4	7.7
Joint Decipherment	18.9	18.0	9.8

Table 5: Decipherment for Malagasy Translation

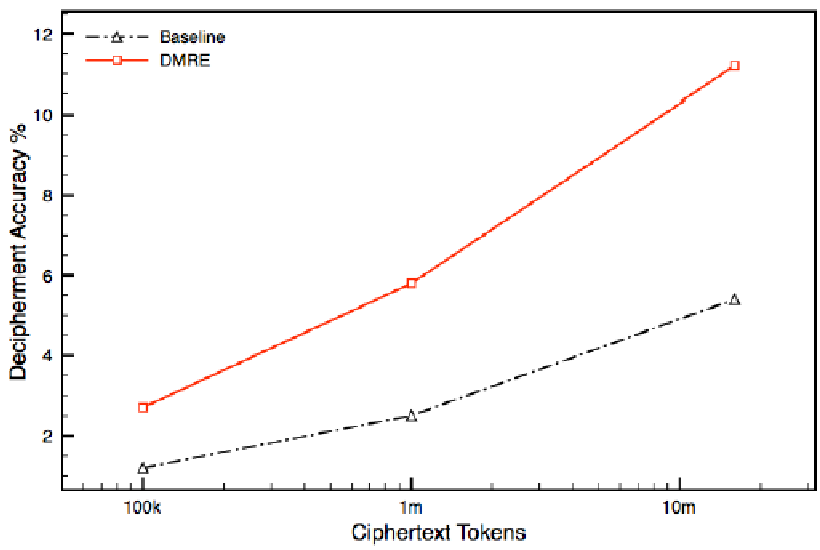

Figure 1: Our new method of decipherment with word embeddings (red line) yeilds significantly better translation dictionaries when applied to non-parallel Malagasy ("ciphertext") and English ("plaintext") collections.

the benefit of a parallel text. In this work, carried out jointly by ISI and CMU, we introduced a new similarity matrix that relates Malagasy word embeddings with English ones. We use this as an evolving base distribution during decipherment training, obtaining an ultimate word-to-word mapping that is twice as accurate as previously possible (Figure 1). The method also produces gains for Spanish/English, and we have released a generic decipherment toolkit (MonoGiza).

Swahili: We have shown that lexical correspondences induced using models of lexical borrowing can project resources – namely, translations – leading to improved performance in a downstream translation system ([65]). Our solution is depicted in Figure 2. Given an out-of-vocabulary (OOV) word in resource-poor SwahiliEnglish MT, we plug it into the Swahili Arabic borrowing system that

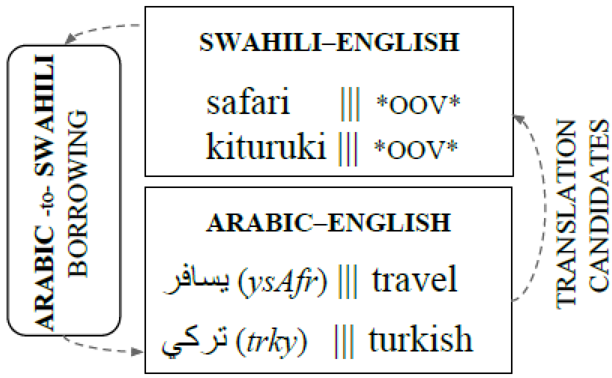

Figure 2: MT with OOV candidates selected from our model of borrowing

identifies the list of plausible Arabic donor words. Then, using resource-rich ArabicEnglish MT, we translate the Arabic donor words to English. Finally, we integrate translation candidates in the resource-poor system as synthetic phrases, a technique for improving translation that we developed in previous years. To let the translation model learn whether to trust these phrases, the translation options obtained from the borrowing model are augmented with a boolean translation feature indicating that the phrase was generated externally, plus semantic and phonetic features corresponding to properties of the donorloan words relation.

Figure 3 summarizes SwahiliEnglish MT experiments. SwahiliEnglish MT performance is improved by up to +1.6 BLEU when we augment it with translated OOV loanwords leveraged from the ArabicSwahili borrowing and then ArabicEnglish MT. The contribution of the borrowing dictionaries is +0.61.1 BLEU, and phonetic and semantic features contribute additional half BLEU. More importantly, upper bound results show that the system can be improved more substantially with better dictionaries of OOV loanwords. This result confirms that OOV borrowed words is an important type of OOVs, and with proper modeling it has the potential to improve translation by a large margin.

	4K	8K	14K
Baseline	13.2	15.1	17.1
+ Transliteration OOVs	13.4	15.3	17.2
+ Loan OOVs	14.3	15.7	18.2
+ Features	**14.8**	**16.4**	**18.4**
Upper bound loan	18.9	19.1	20.7
Upper bound all OOVs	19.2	20.4	21.1

Figure 3: MT with OOV candidates selected from our model of borrowing

3 Text Analysis (TA) and Natural Language Processing (NLP) Methods

3.1 Graph Transduction

Weighted finite-state acceptors and transducers are a critical technology for NLP and speech systems. They flexibly capture many kinds of left-to-right substitutions, and simple transducers can be composed into more complex cascades. These automata are also trainable from data. However, they are unable to capture long-range syntactic dependencies that characterize natural language. Fortunately, tree acceptors and transducers address this weakness. Tree automata have been profitably used in syntax-based machine translation (MT) systems. Still, strings and trees are both weak at representing linguistic structure involving semantics and reference ("who did what to who").

Graph-based feature structures provide an attractive, well-studied, standard semantic format. Over the course of this project, we developed algorithms for probabilistic acceptors and transducers over graph feature structures, and we laid down a formal foundation for semantics-based statistical machine translation. This project made use of the first large-scale, whole-sentence semantic annotation of English sentences using Abstract Meaning Representation (AMR) [5]. The AMR database contains tens of thousands of English sentences pairs with manually-constructed meaning graphs. This provides both testing and training material for automata that can transduce English strings into semantic graphs, and vice-versa.

We pursued two types of weighted graph automata, examined their theoretical properties, and built algorithms that implement generic automata operations. (1) DAG acceptors and DAG-to-tree transducers modeled after [43]. We modified the formalisms to fit edge-labeled semantic graphs, and we built algorithms for membership checking, transduction, and k-best extraction. We packaged these in a toolkit called DAGGER [57, 56]. (2) Hyperedge replacement grammars (HRG),

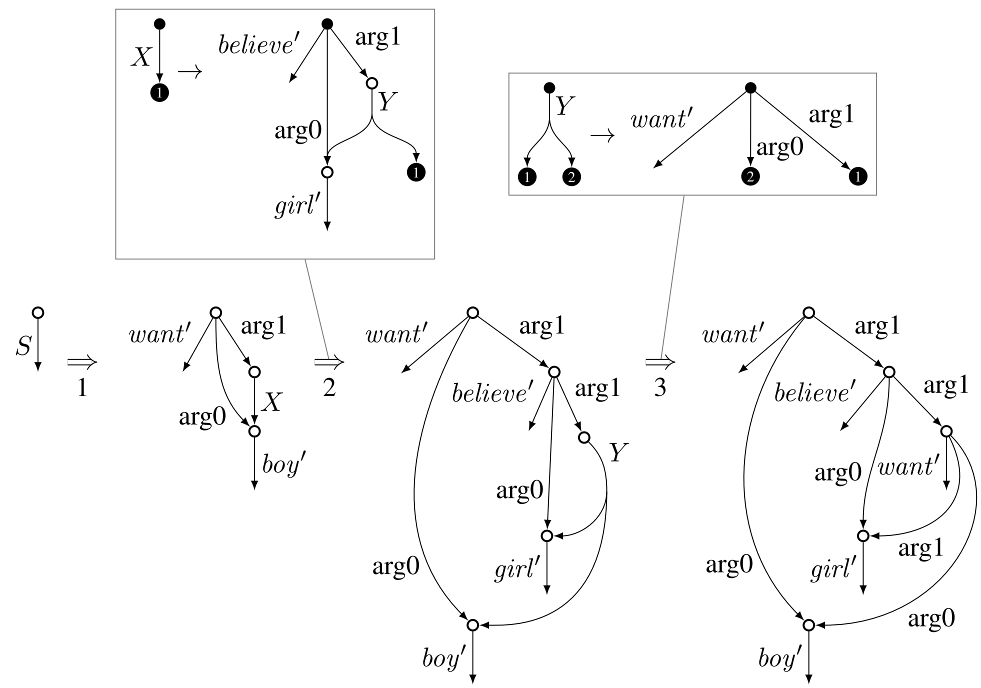

Figure 4: Derivation of a semantic graph representing the meaning of "The boy wants the girl to believe that he wants her," using rules from a hyperedge replacement grammar (HRG).

modeled after [20]. We developed a novel algorithm [12] for analyzing an input graph G according to a HRG, i.e., extracting all derivations (if any) of G and packing them into an efficient data structure. This algorithm runs in time O($(2^{dn})^{k+1}$), where d and n are properties of G, and k is a property of the HRG. This algorithm forms the basis of efficient transduction using synchronous HRG (SHRG). We also exploited the context-free derivation structure of HRG to get efficient algorithms for k-best extraction and expectation-maximization (EM) training. HRGs are not closed under intersection, which distinguishes them from string and tree acceptors.

Figure 4 illustrates the derivation of a semantic graph representing the meaning (predicate-argument and co-referential structure) of a complex sentence in three steps.

Using these graph parsing algorithms we developed a method for semantics-based MT and implemented a prototype system [41]. The mapping from source sentence to graph-based meaning representation and the meaning representation into English are learned from annotated data. The source language is mapped into a target language semantic graph meaning representation which is then transformed into a parse tree and finally a parse string.

We also carried out an empirical comparison of graph grammars and automata [7]. We investigated formalisms for capturing the relation between semantic graphs and English strings. Because it has not been clear which formalisms are a good fit for natural language data—in particular, for describing how semantic graphs represent English pronouns, zero pronouns, reflexives, passives, nominalizations, etc.—we introduced a data set that focuses on these problems, consisting of 10,000 synthetically-generated graph/sentence pairs. We built grammars to capture the graph/string relation in this data, and we evaluated those grammars for conciseness and accuracy.

Beyond algorithms and scientific reports, this project produced the graph-transducer software toolkit Bolinas,[1] and it also brought together computing theory researchers from the US and Europe, leading to intense renewed research on graph languages. In March 2015, a Dagstuhl Seminar held in Germany (*Formal Models of Graph Transformation in Natural Language Processing*),[2] brought the PIs together with 30 experts in the field and resulted in a number of new collaborations.

3.2 Morphological Parsing and Modeling

Modeling Syntactic Context Improves Unsupervised Morphological Segmentation [44] The connection between part-of-speech (POS) categories and morphological properties is well-documented in linguistics but underutilized in text processing systems. This paper proposes a novel model for morphological segmentation that is driven by this connection. Our model learns that words with common affixes are likely to be in the same syntactic category and uses learned syntactic categories to refine the segmentation boundaries of words. Our results demonstrate that incorporating POS categorization yields substantial performance gains on morphological segmentation of Arabic. This work relates to MURI in improving accuracy of TA tools and possibly reducing the amount of data required to achieve high accuracy.

Unsupervised Morphological Analyzer Most state-of-the-art systems today produce morphological analysis based only on orthographic patterns. In contrast, we propose a model for unsupervised morphological analysis that integrates orthographic and semantic views of words. We model word formation in terms of mor-

[1] www.isi.edu/licensed-sw/bolinas

[2] www.dagstuhl.de/de/programm/kalender/semhp/?semnr=15122

Language	Correct Segmentations		Incorrect Segmentations		
	Word	**Segmentation**	**Word**	**Predicted**	**Correct**
English	salvoes	salvo-es	contempt	con-tempt	contempt
	negotiations	negotiat-ion-s	sterilizing	steriliz-ing	steril-iz-ing
	telephotograph	tele-photo-graph	desolating	desolating	desolat-ing
	unequivocally	un-equivocal-ly	storerooms	storeroom-s	store-room-s
	carsickness's	car-sick-ness-'s	tattlers	tattler-s	tattl-er-s
Turkish	moderni	modern-i	mektuplaşmalar	mektuplaşma-lar	mektup-laş-ma-lar
	teknolojideki	teknoloji-de-ki	gelecektiniz	gelecek-tiniz	gel-ecek-ti-niz
	burasıydı	bura-sı-ydı	aynalardan	ayna-lar-da-n	ayna-lar-dan
	çizgisine	çiz-gi-si-ne	uyuduğunuzu	uyudu-ğu-nuzu	uyu-duğ-unuz-u
	değişiklikte	değişik-lik-te	dirseğe	dirseğe	dirseğ-e
Arabic	sy$Ark	s-y-$Ark	wryfAldw	w-ry-fAldw	w-ryfAldw
	nyqwsyA	nyqwsyA	bHlwlhA	b-Hlwl-h-A	b-Hlwl-hA
	AlmTrwHp	Al-mTrwH-p	jnwby	jnwb-y	jnwby
	ytEAmlwA	y-tEAml-wA	wbAyrn	w-bAyr-n	w-bAyrn
	lAtnZr	lA-t-nZr	rknyp	rknyp	rkny-p

Figure 5: Examples of correct and incorrect segmentations produced by our unsupervised morphological analyzer on three languages. Correct segmentations are taken directly from gold MorphoChallenge data.

phological chains, from base words to the observed words, breaking the chains into parent-child relations. For instance, given a word playfully, the corresponding chain is play $\rightarrow$ playful $\rightarrow$ playfully. The word play is a base form of this derivation as it cannot be reduced any further. Individual derivations are obtained by adding a morpheme (ex. ful) to a parent word (ex. play). This addition may be implemented via a simple concatenation, or it may involve transformations. At every step of the chain, the model aims to find a parent-child pair (ex. play/playful) such that the parent also constitutes a valid entry in the lexicon. This allows the model to directly compare the semantic similarity of the parent-child pair, while also considering the orthographic properties of the morphemic combination.

We model each step of a morphological chain by means of a log-linear model that enables us to incorporate a wide range of features. At the semantic level, we consider the relatedness between two words using the corresponding vector embeddings. At the orthographic level, features capture whether the words in the chain actually occur in the corpus, how affixes are reused, as well as how the words are altered during the addition of morphemes. We use Contrastive Estimation [62] to efficiently learn this model in an unsupervised manner. Specifically, we require that each word has greater support among its bounded set of candidate parents than an artificially constructed neighboring word would.

Our model consistently matches or outperforms five state-of-the-art systems on Arabic, English and Turkish.

Morphology-Aware Statistical Language Models Swahili, Kinyarwanda, and Malagasy make extensive use of morphological processes to express diverse grammatical features, including tense, voice, aspect, agreement, and evidentiality. The morphological richness of these languages means that traditional surface-level statistical models are ineffective. We then developed a general purpose technique for incorporating these morphological grammars in state-of-the-art nonparametric Bayesian statistical models [10]. Using these techniques, we are able to (1) infer likely missing elements in the lexicon that human editors could correct, (2) estimate the most likely analysis of a word form in context, (3) achieve state-of-the-art language modeling results in a variety morphologically rich languages, and (4) achieve state-of-the-art word alignment results between English and morphologically rich languages.

We additionally developed state of the art analyzers for Swahili and Kinyarwanda, obtaining token-level coverage of 85% and 97%. While this is a high rate of coverage, the proper analysis of a word is ambiguous in context, so even with these coverage rates, our statistical techniques are crucial.

3.3 Unsupervised NLP models

We designed, implemented and tested a range of unsupervised models for NLP tasks. These include:

- A bilingual part of speech model based on feature-rich Markov random fields [11].
- A method for transferring information in supervised models for one or more resource-rich languages to an unsupervised learner for a resource-poor language, tested on part-of-speech tagging and parsing [14].
- A model for discovering multi-word, gappy expressions in monolingual and bilingual text, evaluated within a translation system [36].
- A model for word alignment based on feature-rich conditional random fields [24].

Light Semantic Annotations Improve Unsupervised Semantic Annotations [51] We developed a method for dependency grammar induction that utilizes sparse annotations of semantic relations. This induction set-up is attractive because such annotations provide useful clues about the underlying syntactic structure, and they are readily available in many domains (e.g., info-boxes and HTML markup). Our method is based on the intuition that syntactic realizations of the same semantic predicate exhibit some degree of consistency. We incorporate this intuition in a directed graphical model that tightly links the syntactic and semantic structures. This design enables us to exploit syntactic regularities while still allowing for variations. Another strength of the model lies in its ability to capture non-local dependency relations. Our results demonstrate that even a small amount of semantic annotations greatly improves the accuracy of learned dependencies when tested on both in-domain and out-of-domain texts. This work was evaluated on English, but pertains to the improvement of parsing accuracy with less training data.

Unsupervised Chunking We completed modeling and experiments for learning, from raw text, to identify low-level syntactic constituents, an unsupervised version of text chunking. For example, in a sentence like the man gave a big book to the boy in London. the algorithm would ideally find a segmentation such as [the man] gave [a big book] to [the boy] [in London]. We showed that addressing this task directly, using probabilistic finite-state methods, produces better results than relying on the local predictions of state of the art unsupervised constituency parsers. Furthermore, by running the same model in a cascade, we were able to identify higher-level constituents, allowing us to produce (full-sentence) constituent structures. Doing so outperforms previous work by a wide margin in unlabeled parsing evaluation scores for English, German and Chinese. This work is reported in [55]. Now that we have syntactic annotations available for Kinyarwanda and Malagasy, we can begin to test the models on the focus languages.

3.4 Part-of-Speech Tagging

Rapid development of NLP in a new domain: In collaboration with many others, we considered part-of-speech tagging in social media (specifically, Twitter). We modified the POS lexicon, annotated data, and constructed new features appropriate for the domain. This exercise helped us see the viability of rapid adaptation of NLP tools for new text domains [35].

We introduced an improved method for learning part-of-speech taggers from a human-supplied tag dictionary [31]. Previous work has generally assumed an experimentally vexed setup that masked many issues that arise when using information about types (e.g., which parts-of-speech some set of words in the languages can associate with). The technique greatly improves performance for a more realistic scenario in which the tag dictionary is obtained from a source that is disjoint from the raw data used to learn a Hidden Markov model (and from the data used to evaluate the method), shown for English and Italian.

An automatic method for mapping language-specific part-of-speech tags to

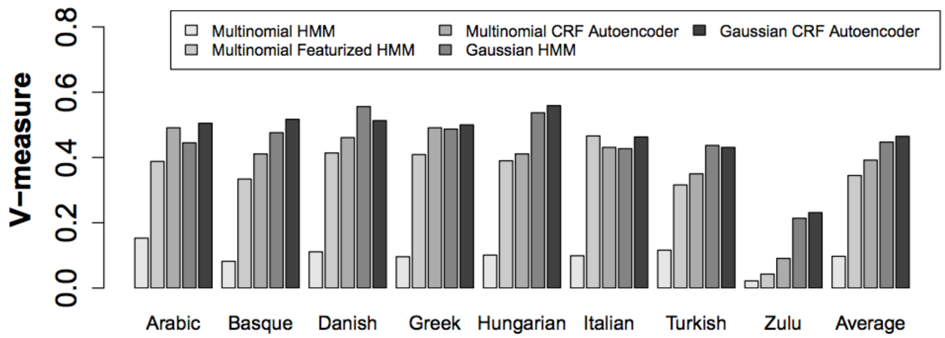

Figure 6: Models which use standard skip-gram word embeddings (i.e., Gaussian HMM and Gaussian CRF Autoencoder) outperform all baselines on average across languages

a set of universal tags: [70]. This unified representation plays a crucial role in cross-lingual syntactic transfer of multilingual dependency parsers (see above). Until now, however, such conversion schemes have been created manually. Our central hypothesis is that a valid mapping yields POS annotations with coherent linguistic properties which are consistent across source and target languages. We encode this intuition in an objective function that captures a range of distributional and typological characteristics of the derived mapping. Given the exponential size of the mapping space, we propose a novel method for optimizing over soft mappings, and use entropy regularization to drive those towards hard mappings. Our results demonstrate that automatically induced mappings rival the quality of their manually designed counterparts when evaluated in the context of multilingual parsing.

Low-Resource Part-of-Speech Tagging A single word may have multiple parts of speech. For example, in English, the word book may refer to the printed volume on the shelf, or may be a verb referring to making a travel reservation. Thus, annotators have traditionally focused on labeling words with their part-of-speech in context. Our year 3 work has shown that type-level annotation, in which annotators list all possible parts of speech for different word types, independent of context. To further improve performance, we infer POS dictionaries for unlabeled word types using a graph-based semi-supervised learning algorithm. In sum, our approach is far more effective (in terms of effort required and final quality) for obtaining high quality part-of-speech taggers in new languages and domains [32, 34]. We demonstrate this by creating POS taggers in English using experienced and novice annotators, as well as for the focus languages, Kinyarwanda and Malagasy.

Gaussian Reconstruction in CRF Autoencoders In [3] we introduced the conditional random field (CRF) autoencoder model for unsupervised learning of part-of-speech tags. One limitation of this model has been the use of a multinomial distribution to reconstruct words given a POS tag, which prevents related words from sharing model parameters. In [47] we use a multivariate Gaussian distribution to reconstruct pretrained embeddings given a POS tag. Results show a significant improvement over the original CRF autoencoder model.

3.5 Named Entity Recognition

Transliteration of Named Entities Proper names are often translated across languages using a process that largely respects the pronunciation and spelling of the name in the original language, called transliteration. We introduced a new model for named entity transliteration generation. The model learns from noisy transliterations collected from the web, which are the only training data available for many low-resource languages. We entered this model into the ACL 2012 Named Entity Workshop shared task [2]. The technique achieved competitive performance on the Arabic-English task and can be adapted to any language pair for which noisy annotated examples are available.

3.6 Representation Learning

Sparse Overcomplete Word Vector Representations Current distributed representations of words show little resemblance to theories of lexical semantics. The former are dense and uninterpretable, the latter largely based on familiar, discrete classes (e.g., supersenses) and relations (e.g., synonymy and hypernymy). We have developed methods that transform word vectors into sparse (and optionally binary) vectors [27]. The resulting representations are more similar to the interpretable features typically used in NLP, though they are discovered automatically from raw corpora. Because the vectors are highly sparse, they are computationally easy to work with. Using a number of state-of-the-art word vectors as input, we find consistent benefits of our method on a suite of standard benchmark evaluation tasks.

We also evaluate our word vectors in a word intrusion experiment with humans and find that our sparse vectors are more interpretable than the original vectors. We anticipate that sparse, binary vectors can play an important role as features in statistical NLP models, which still rely predominantly on discrete, sparse features whose interpretability enables error analysis and continued development.

Vectors		SimLex Corr.	Senti. Acc.	TREC Acc.	Sports Acc.	Comp. Acc.	Relig. Acc.	NP Acc.	Average
Glove	Initial	36.9	77.7	76.2	95.9	79.7	86.7	77.9	76.2
	Sparse	38.9	**81.4**	**81.5**	**96.3**	**87.0**	**88.8**	**82.3**	**79.4**
SG	Initial	**43.6**	81.5	77.8	97.1	80.2	85.9	80.1	78.0
	Sparse	41.7	**82.7**	**81.2**	**98.2**	**84.5**	**86.5**	**81.6**	**79.4**
GC	Initial	9.7	68.3	64.6	75.1	60.5	76.0	79.4	61.9
	Sparse	**12.0**	**73.3**	**77.6**	**77.0**	**68.3**	**81.0**	**81.2**	**67.2**

Figure 7: Performance comparison of transformed sparse vectors and initial vectors

Hierarchical Tensor-based Feature Representations Accurate multilingual transfer parsing typically relies on careful feature engineering. These features range from standard arc features used in monolingual parsers to typological properties needed to guide cross-lingual sharing (e.g., verb-subject ordering preference). Tensor-based models are an appealing alternative to manual feature design. These models automatically induce a compact feature representation by factorizing a tensor constructed from atomic features.

We propose a hierarchical tensor-based approach for this task. This approach induces a compact feature representation by combining atomic features. However, unlike traditional tensor models, it enables us to incorporate prior knowledge about desired feature interactions, eliminating invalid feature combinations. To this end, we use a hierarchical structure that uses intermediate embeddings to capture desired feature combinations. At the bottom level of the hierarchy, the model constructs combinations of atomic features, generating intermediate embeddings that represent the legitimate feature groupings. For instance, these groupings will not combine the verb-subject ordering feature and the POS head feature. At higher levels of the hierarchy, the model combines these embeddings as well as the expert-defined typological features over the same atomic features.

The hierarchical tensor is thereby able to capture the interaction between features at various subsets of atomic features. Algebraically, the hierarchical tensor is equivalent to the sum of traditional tensors with shared components. Thus, we can use standard online algorithms for optimizing the low-rank hierarchical tensor.

In both unsupervised and semi-supervised transfer scenarios, our hierarchical tensor consistently improves UAS (unlabeled attachment score) and LAS (labeled attachment score) over state-of- the-art multilingual transfer parsers and the traditional tensor model across 10 different languages.

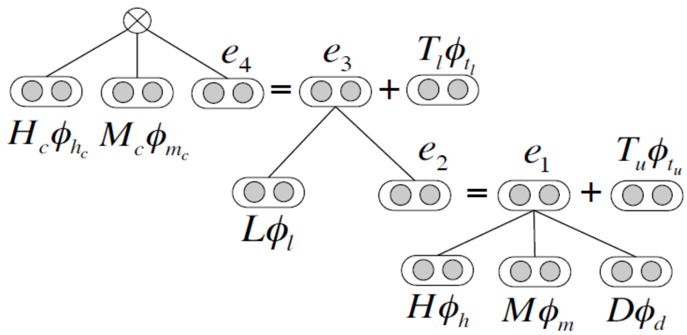

Figure 8: Hierarchical Tensors

3.7 Dependency and Constituency Parsing

Multilingual Dependency Parsing We developed a novel algorithm for multilingual dependency parsing that uses annotations from a diverse set of source languages to parse a new unannotated language [52]. Our motivation is to broaden the advantages of multilingual learning to languages that exhibit significant differences from existing resource-rich languages. The algorithm learns which aspects of the source languages are relevant for the target language and ties model parameters accordingly. The model factorizes generation of a dependency tree into two processes: selection of syntactic dependents and their ordering. Being largely language-independent, the selection component is learned in a supervised fashion from all the training languages. In contrast, the ordering decisions are only influenced by languages with similar properties. We systematically model this cross-lingual sharing using typological features. We evaluate our selective sharing model on 17 languages from multiple language families. In our experiments, the model consistently outperforms a state-of-the-art multilingual parser, including our multilingual guidance technique [14].

Unsupervised Dependency Parsing Orthogonally, we introduced a new initializer for unsupervised dependency parsing based on convex optimization of a simpler unsupervised dependency model [38]. The technique allows incorporation of some kinds of linguistic knowledge (e.g., preferences for the types of words likely to be heads of the sentence). On average, the model leads to attachment accuracy gains over the classic initializer.

Low-resource Natural Language Parsing In addition to the supervised dependency parsers we have developed and released for Kinyarwanda and Malagasy (discussed below in the tools section), we have developed new techniques for learning parsing more effectively with very few annotations in the language of interest, in a variety of formalisms (CFG, CCG, and HPSG). A semi-supervised method was developed that combines a coarse unlexicalized model trained on a small amount of manual annotations with a refined lexicalized model trained on large amounts of automatically annotated data. The combination is then used to re-annotate the unlabeled data and the whole process is repeated. The performance improvement achieved by this technique is roughly equivalent to doubling the amount of manual annotations. Furthermore, the coarse model performance when trained on small amounts of data is significantly better than the standard supervised parsers designed to learn from large annotated corpora [68].

Automata for Linguistic Analysis [12] developed a tractable, novel graph parsing algorithm and other efficient procedures for generating, recognizing, and transforming directed acyclic graphs. This algorithm provides the formal machinery for modeling translation using semantic graphs, which derive predicate-argument relations and co-reference from the recursive application of rules that "add to" partially constructed meanings until the full proposition has been expressed.

Supertagging with Weak Supervision A subteam of collaborators from the University of Texas and CMU designed, implemented, and tested a Bayesian formulation for weakly-supervised learning of a Combinatory Categorial Grammar (CCG) supertagger [33]. The approach is based on a hidden Markov model and assumes supervision in the form of a tag dictionary. Our prior encourages the use of cross-linguistically common category structures as well as transitions between tags that can combine locally according to CCG's combinators. This prior is theoretically appealing, since it is motivated by language-independent, universal properties of the CCG formalism.

Empirically, we show that it yields substantial improvements over previous work that used similar biases to initialize an EM-based learner (Baldridge, 2008). Additional gains are obtained by further shaping the prior with corpus-specific information that is extracted automatically from raw text and a tag dictionary.

Latent PCFG Parsing for Low-resource Languages We have adapted Gibbs samplers for PCFGs to work with latent annotations and showed it works well across a wide variety of languages, including English, Chinese, Italian, Malagasy, and Kinyarwanda [63], and is competitive with previous approaches. It works especially well for low-resource languages, beating both standard Bikel parser and PCFGs with latent annotations estimated by variational methods.

We have also extended the Gibbs sampler to estimate dependency grammars from no training data and partial training data (as given by GFL). We are now annotating GFL for English, Portuguese, Chinese and Kinyarwanda to provide training data for this use case. This is a collaboration between Texas and CMU.

Sampling-based Inference We have developed a sampling-based inference algorithm for dependency parsing that can handle an arbitrary scoring function. Starting with an initial candidate tree, our inference procedure climbs the scoring function in small (cheap) stochastic steps towards a high scoring parse. The proposal distribution over the moves is derived from the scoring function itself. Because the steps are small, the complexity of the scoring function has limited impact on the computational cost of the procedure. We explore two alternative proposal distributions. Our first strategy is akin to Gibbs sampling and samples a new head for each word in the sentence, modifying one arc at a time. The second strategy relies on a provably correct sampler for first-order scores, and uses it within a Metropolis-Hastings algorithm for general scoring functions. It turns out that the latter optimizes the score more efficiently than the former.

The benefits of sampling-based learning go beyond stand-alone parsing. For instance, we can use the framework to correct preprocessing mistakes in features such as part-of-speech (POS) tags. In this case, we combine the scoring function for trees with a stand-alone tagging model. When proposing a small move, i.e., sampling a head of the word, we can also jointly sample its POS tag from a set of alternatives provided by the tagger. As a result, the selected tag is influenced by a broad syntactic context above and beyond the initial tagging model and is directly optimized to improve parsing performance. Our joint parsing-tagging model provides an alternative to the widely-adopted pipeline setup. This functionality is particularly significant for low-resource languages where POS taggers are expected to be noisy [69].

Tensor-based Dependency Representation We have developed a low-rank factorization model for scoring dependency arcs and have applied it to first- and third-order dependency parsing. Accurate scoring of syntactic structures such as head-modifier arcs in dependency parsing typically requires rich, high-dimensional feature representations. A small subset of such features is often selected manually. This is problematic when features lack clear linguistic meaning as in embeddings or when the information is blended across features. We use tensors to map high-dimensional feature vectors into low-dimensional representations. We explicitly maintain the parameters as a low-rank tensor to obtain low dimensional representations of words in their syntactic roles, and to leverage modularity in the tensor

Train			Test		
P	R	F_1	P	R	F_1
.87	.85	.86	.82	.72	.76

Table 6: AMR concept identification performance.

for easy training with online algorithms [45].

Semantic Dependency Parsing We designed, implemented, and tested an approach to semantic dependency parsing. A semantic dependency parse is represented as a directed graph where (as in syntactic dependency parsing) words are vertices. Edges represent semantic relationships between words, and the collection of edges need not form a tree. Our approach is based on graph-based methods for syntactic dependency parsing. It was evaluated in SemEval 2014 Task 8, a competition that considered three different formalisms for semantic dependencies, and an English dataset for each one. Our approach came in second place overall. The first-place system used a similar approach but with higher- order features. It is worth noting that their approach was based on TurboParser, a tool produced in past work in collaboration with members of this team.

Abstract Meaning Representation Parsing and Generation A more powerful, and more theoretically compelling formalism is the Abstract Meaning Representation or AMR [6]. We designed, implemented, and evaluated the first parsing algorithm specifically for AMR [30]. The method is based on a novel algorithm for finding a maximum spanning, connected subgraph of a graph, embedded within a Lagrangian relaxation of an optimization problem imposing linguistically inspired determinism constraints. The design of the system, in the framework of structured prediction, allows future incorporation of additional features and constraints, and may be applied to other semantic formalisms. The implemented system, JAMR, has been made available as open-source code and is already in used by several research groups around the world. We evaluated the system on a held-out 2,100-sentence test set; results are shown in Tables 6 and 7. For concept identification (the first stage of analysis), we compare labeled spans against automatically induced spans from our aligner. For parsing, we report the Smatch version 1.0 scores [8].

One of the goals of AMR is to be suitable as an intermediate semantic representation for tasks such as machine translation and summarization. To enable this use case, we have designed and built the first method for generation of natural language

	Train			Test		
Concepts	P	R	F_1	P	R	F_1
Gold	.85	.95	.90	.76	.84	.80
Automatic	.69	.78	.73	.52	.66	.58

Table 7: AMR parser performance.

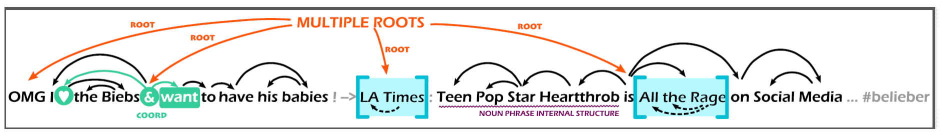

Figure 9: Tweeboparser parse tree for a constructed tweet

from AMR [29]. Our approach to generation from AMR is to convert the AMR graph to a tree, and then to generate from the tree using a tree-to-string transducer. The rules for the tree-to-string transducer are learned from the AMR corpus, and the system is trained discriminatively with features that include a language model. We evaluated the system using BLEU [54] and achieved 21 single-reference BLEU on the standard test split (LDC2014T12).

Joint Modeling using Randomized Greedy Inference We introduce a new approach for joint segmentation, POS tagging and dependency parsing. While joint modeling of these tasks addresses the issue of error propagation inherent in traditional pipeline architectures, it also complicates the inference task. Past research has addressed this challenge by placing constraints on the scoring function. In contrast, we propose an approach that can handle arbitrarily complex scoring functions. Specifically, we employ a randomized greedy algorithm that jointly predicts segmentations, POS tags and dependency trees. Moreover, this architecture readily handles different segmentation tasks, such as morphological segmentation for Arabic and word segmentation for Chinese. The joint model outperforms the state-of-the-art systems on three datasets, obtaining 2.1% TedEval absolute gain against the best published results in the 2013 SPMRL shared task.

Tweeboparser Parser for Tweets We designed an approach for statistical dependency parsing of English tweets. Tweets are interesting because they present text data in a form which is used on a daily basis by users all over the world. It is in contrast with the formal text in the Wall Street Journal, which has been used to generate the Penn Treebank, which serves as the training data for almost all English parsers to date. However, tweets present a new challenge with their non-standard use of language, primarily attributed to a limitation of 140 characters per tweet. This, in addition to Twitter-specific constructs such as hashtags and retweets, cause a standard statistical parser to fail to provide a reasonable syntactic analysis of the tweet.

Our approach involved a new annotation scheme for English tweets, based on the Graph Fragment Language (GFL) scheme [60]. Most of our annotators were graduate students from CMU, and received very brief training. For the statistical parser, we used a modified approach that allows more flexible feature engineering for tweets. Our parser is an extension of TurboParser, which was built in previous collaboration with members of this group. We obtained nearly 81.0% unlabeled attachment accuracy under this approach.

Transition-based Dependency Parsing with S-LSTMs In [22] we proposed a technique for learning representations of parser states in transition-based dependency parsers. Our primary innovation is a new control structure for sequence-to-sequence neural networks - the stack long short-term memory model (S-LSTM). Like the conventional stack data structures used in transition-based parsing, elements can be pushed to or popped from the top of the stack in constant time, but, in addition, an LSTM maintains a continuous space embedding of the stack contents. This lets us formulate an efficient parsing model that captures three facets of a parser's state: (i) unbounded look-ahead into the buffer of incoming words, (ii) the complete history of actions taken by the parser, and (iii) the complete contents of the stack of partially built tree fragments, including their internal structures. Standard backpropagation techniques are used for training and yield state-of-the-art parsing performance on the English Penn Treebank.

Cross-lingual Dependency Parsing [4] A popular approach to develop dependency parsers for low-resource languages is to train a "delexicalized" parser on the treebanks in a high-resource source language. This approach traditionally suffered from the lack of lexical features, which are known to outperform. We use a novel method for estimating multilingual embeddings to recover some lexical information which can be transferred across languages. We also train the parser on treebanks available in several source languages to leverage the syntactic diversity exhibited in those treebanks, while using typological properties of each language as extra input to the parser. Experiments on v1.0 of the universal dependency treebanks (Figure 10) show that this approach can effectively combine supervision from multiple source languages, without specifying the target language in advance.

UAS	target language										(avg)
source language ↓	cs	de	en	es	fi	fr	ga	hu	it	sv	
cs	-	62.5	54.9	62.8	46.7	59.5	54.8	52.9	65.8	51.5	*56.8*
de	58.8	-	57.9	60.9	47.7	58.2	51.9	54.5	62.2	53.1	*56.1*
en	49.3	60.3	-	63.7	49.3	61.7	49.7	47.7	63.9	55.5	*55.7*
es	59.3	58.6	60.1	-	43.8	71.4	55.2	46.7	74.4	54.1	*58.2*
fi	46.3	52.5	50.6	45.1	-	43.0	34.9	48.6	44.8	43.3	*45.4*
fr	54.1	57.2	61.6	69.4	42.5	-	55.6	46.5	72.5	52.1	*56.8*
ga	31.6	34.9	33.8	49.1	22.1	45.8	-	22.9	52.9	33.0	*36.2*
hu	41.9	52.6	46.4	40.9	48.8	39.6	31.5	-	38.4	40.2	*42.3*
it	56.4	53.1	53.4	70.7	42.0	70.0	52.6	41.9	-	52.2	*54.7*
sv	48.0	53.9	50.6	54.7	44.7	54.3	50.5	49.3	58.1	-	*51.6*
`Most similar`	46.3	53.9	50.6	**70.7**	46.7	**71.4**	54.8	48.6	**74.4**	53.1	57.0
`Multiple`	**52.2**	**61.0**	**61.5**	69.1	**51.3**	68.8	**58.0**	**54.3**	67.9	**57.4**	**60.1**

LAS	target language										(avg)
source language ↓	cs	de	en	es	fi	fr	ga	hu	it	sv	
`Most similar`	35.1	37.7	35.4	**61.8**	35.6	**63.2**	27.2	29.8	**66.0**	42.4	43.4
`Multiple`	**40.2**	**51.2**	**52.4**	60.3	**36.8**	59.7	**41.6**	**40.8**	59.6	**45.9**	**48.8**

Figure 10: Cross-lingual dependency parsing with one vs. multiple source languages. Top table: Unlabeled Attachment Score; Bottom Table, Labeled Attachment Score

4 Machine Translation

We explored a range of ideas for improving statistical machine translation systems that we expect will be applicable in resource-poor settings.

- Implemented a hierarchical phrase-based German-English translation system that was ranked second in the SMT competition (after Google). This system incorporated a discriminative German parsing model trained only on parallel data (i.e., no treebank), long-range unsupervised class-based language models, and a latent variable compound splitter [25].

- Designed, implemented, and tested a new translation model based on dependencies over phrases. Experimented with unsupervised parsing, reporting only very small performance decreases when moving from supervised to unsupervised parsers [37].

- Explored methodology for testing hypotheses about translation systems, leading to practical recommendations for researchers in the field [13].

4.1 Crowdsourcing for Low Resource MT

We developed means of crowd-sourcing data and annotations to create and improve low-resource machine translation systems. A sequence of experiments conducted on several language pairs concluded the following:

- Bilinguals with no linguistic expertise are able to: provide translations (more below); provide alignments at the word level with reasonable reliability and replicability; provide judgments as to the quality of a machine-generated translation; provide phrasal and lexical level cross-language equivalents.

- Compared with expert (linguist or translator) translations, non-expert crowd-sourced translations may not be of the same quality, but obtaining three or more different translations of the same source sentence from non-experts improves an MT system more than obtaining a single expert translation, and can be far more cost effective. Hence, the best use of linguistic expertise is not generating translations or rating them, but focusing on tasks that non-expert bilinguals are un- able to do, such as linguistic annotations (syntactic and semantic) and linguistic rule writing.

- A few lessons learned about crowd-sourcing:
  - Instructions must be crystal clear and brief. Crowds do not read beyond the first paragraph.
  - Always pilot a crowd-sourcing task. Crowds will try to bypass anything difficult in the tasks.
  - Always have overlapping assignments to compare results among different workers as validation.

4.2 Modeling

We introduced of a nonparametric prior over synchronous context free grammars (SCFGs) and a Gibbs sampling algorithm for use with the prior, so as to permit inference of translation grammars from sentence- aligned parallel data without constraining the models using word alignments [26].

We developed models that use hand-written morphological analyzers/generators (discussed above) to improve n-gram language models (which are a key component in MT systems). Our preliminary results show that, using only an incomplete analyzer/generator, we can obtain a 40% reduction in model perplexity over a state-of-the-art baseline in Malagasy.

We added latent variables to phrase-based machine translation, with the intention of clustering phrases with the same meaning but different surface forms. Successfully capturing this phenomenon might allow us to estimate translation model parameters more reliably in low-data scenarios. Experiments indicate that this approach does not lead to improvements in translation quality against a strong phrase-based baseline without latent variables.

4.3 Parameter Estimation and Feature Selection

We developed a new learning algorithm for machine translation based on direct minimization of the ramp loss [39]. The algorithm is easy to understand and implement, scales to many features, performs comparably to widely used techniques on benchmark translation tasks, and is more stable across initialization and randomization settings. We also developed a distributed training algorithm for large- scale discriminative training with feature selection [61]. Very large-scale discriminative training for machine translation is challenging in part because of the computational cost of inference, so we developed an algorithm using the MapReduce programming paradigm. This, combined with a procedure for feature selection, means we can scale to orders of magnitude more training data than are standardly used in MT.

4.4 Statistical Semantics-Based Machine Translation

Using the graph parsing algorithms discussed in [12], we have developed a method for semantics-based machine translation and implemented a prototype system [42]. The source language is mapped into a target language semantic graph meaning representation which is then transformed into a parse tree and finally to a target language string.

4.5 Synthetic Translation Options

A key challenge in leveraging knowledge of linguistic processes in machine translation is being able to explain regular "long tail" phenomena without abandoning the ability to deal with irregular (but frequent) forms that are well-handled by traditional statistical approaches that operate on words. To this end, we seek to solve the translation problem in phases, first by predicting translations of words and short phrases uses several specialized modules (targeting, for example, name translation, morphological inflection, verb translation, etc.), and then combining these translation fragments (which we call synthetic translation options) with translation rules learned using from parallel data using statistical techniques into a globally coherent translation. We applied synthetic translation options to address several problems in translation: predicting definite and indefinite articles when translating into English from languages in which definiteness is not lexically indicated [66], generating target language morphology (next section), and incorporating transliteration component into a translation system [1].

We extended work on synthetic phrase creation, a technique for improving translation in low resource languages by generating new entries for the phrase table [67]. In previous years, we focused on morphology and function word creation, this year, we focused on acoustically similar words and showed that we were able to automatically learn alternatives for morphologically related words and improve performance in translation in spoken language conditions, where noisy ASR output is the input to the MT system.

We have also explored using hand-written, high-recall tree-to-string translation rules in pre-translation of the MURI focus languages. This extends notions of pre-reordering of the source language to produce a more target-language-like starting point by adding bracketing demonstratives, deleting articles, and creating morphological alternatives.

4.6 Predicting Target Language Morphology

A major application of our work on synthetic translation options was dealing with the problem of generating appropriately inflected words when translating into morphologically rich languages. Even in high resource scenarios, it is unlikely that the set of possible inflections for each word will be observed in the training data. It is therefore necessary to create new word forms using morphological grammars. In [9], we show how both hand-written morphological grammars and unsupervised grammars can be used to generate contextually appropriate word forms in the target language, based on the source language grammatical context (syntactic dependencies, part of speech sequences, etc.) to improve translation quality.

4.7 Alignment Modeling

In addition to the nonparametric Bayesian morphological model which we use as an alignment model, we developed an effective statistical alignment model that runs 10 times faster than current state-of-the-art systems with the same or better quality [23]. Computing word alignments is a fundamental—and computationally expensive—component of constructing machine translation systems. This innovation enables much more rapid and productive experimentation.

4.8 Discriminative Training

[28] developed a discriminative learning algorithm for machine translation systems that tar- gets mixes of frequently and infrequently occurring features, estimating the sparse features on large amounts of parallel training data. Unlike previous approaches that support learning extremely sparse features, high quality development data continues (which enables better estimation of translation quality) continues to be used to determine the relative contributions of the sparse features in the model along side less-sparse features.

4.9 Shared Task Challenges

NIST Open Machine Translation Evaluation We participated in the NIST Open Machine Translation Evaluation, submitting a Korean-English system. Korean is a morphologically rich language that poses some of the same challenges as our focus languages. The system was based on a hierarchical model with most of the research effort focusing on preprocessing and segmentation. Specifically, we combined a rule-based finite-state morphological analyzer with a nonparametric Bayesian model for morphemes to analyze the Korean text. Performance improvements also came from using a variety of word alignment models and aggregating their output. Details are presented in [26].

WMT 2014 Shared Task We also participated in the 2014 Workshop on Machine Translation shared task, looking at translation from Hindi and German into English. The Hindi task is quite close to the MURI languages since little training data was available, and the language is typologically quite distant to English. We obtained the highest BLEU scores of any of the participants using models that made use of syntactic pre-reordering, synthetic phrase creation, and transliteration techniques that were developed in the MURI project [49].

5 Decipherment

5.1 Decipherment for Machine Translation

Statistical machine translation can perform quite well when trained on large parallel texts. In many languages and domains—in particular those studied in this MURI project—such texts do not exist. We therefore carried out a research program in how to mathematically extract translation knowledge from non-parallel texts. To support the work, we collected 15 million words of Malagasy text from the web. Our approach was to treat Malagasy as a "code" for English, and decipher it. That is, we automatically search for a bilingual Malagasy-English dictionary that, if applied to the Malagasy text, yields good English. Figure 11 illustrates this approach.

Before this MURI project, we had established the theoretical possibility of deciphering foreign languages [58], but this work was confined to very small vocabularies. Our first advance was an algorithm to do large-scale, full-vocabulary decipherment [16]. The algorithm uses *slice sampling* to more efficiently search the space of possible word-for-word decipherments of Malagasy into English. Our first demonstration showed how to improve translation between closely-related languages (French/Spanish) by deciphering a domain-specific French corpus into Spanish. We saw large MT quality improvements of up to 3.8 BLEU points [54].

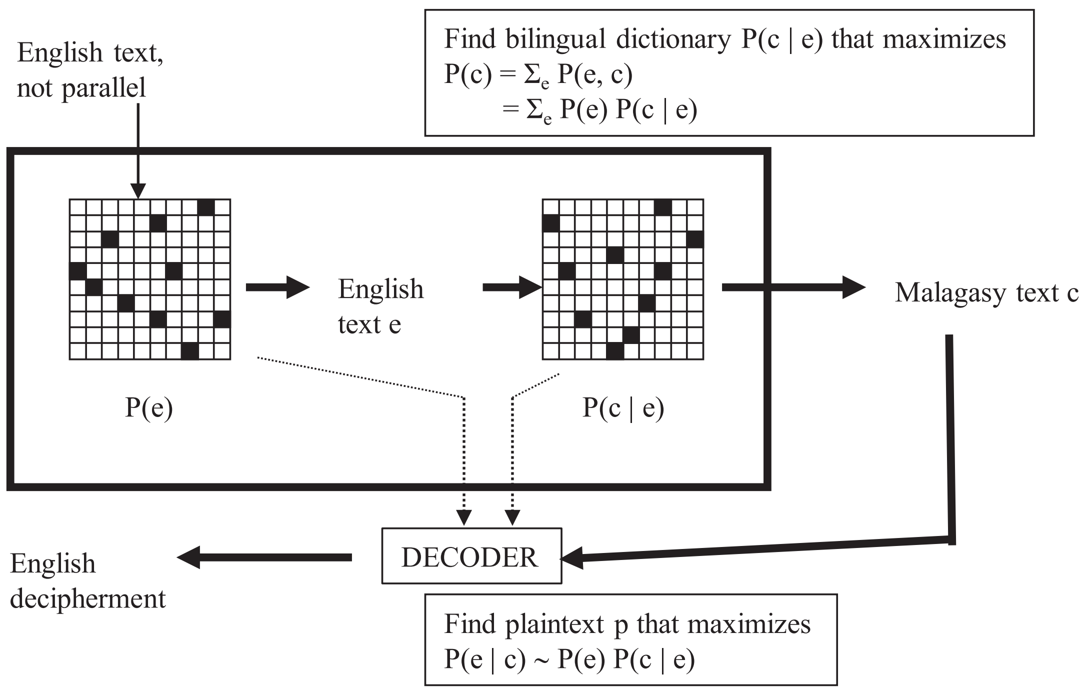

Figure 11: Deciphering Malagasy into English without parallel text. We first train an English language model P(e) (left box), then set up a fully-connected bilingual dictionary P(c | e) (right box). These devices, when composed, generate Malagasy texts according to a distribution. We then observe training text c in Malagasy, and we use it to iteratively revise the bilingual dictionary. Once we have a dictionary, we can translate monolingual Malagasy data into English.

5.2 Syntactic dependency induction

Our next advance [17] introduced syntactic dependency relations to the decipherment process. This allowed us to decipher languages more distantly related than French and Spanish. Word decipherment accuracy rose from 4.2% to 24.6%. We applied the system to translate out-of-vocabulary (OOV) words in a traditional MT system trained on parallel data, resulting in improvements of up to 1.8 Bleu points.

Next, we learned how to profitably combine small parallel text resources with large non-parallel text resources [19]. We used the latter to improve word-level alignment of the parallel text, and we also deciphered it to augment the bilingual dictionary. We tested our new algorithm on Malagasy/English, collecting 2 million words of parallel data, and 15 million words on non-parallel (Malagasy) data. We obtained BLEU score gains of 0.9 and 2.1 on blog and news translation from Malagasy to English.

5.3 Semantic similarity matrices

Finally, in work carried out jointly by ISI and CMU, we introduced a new similarity matrix that relates Malagasy word embeddings with English ones [18]. We used this as an evolving base distribution during decipherment training, obtaining an ultimate word-to-word mapping twice as accurate as previously possible.

We extended our work on noisy-channel decipherment in two additional directions. First, we carried out research on text compression. Bilingual texts contain large amounts of redundancy, in an information-theoretic sense, and understanding that redundancy is critical to translation technology. We introduced a bilingual text compression challenge that asks for the smallest executable that accurately reproduces a large, specific bilingual text, byte-for-byte [71]. We were able to compress a 619.4Mb bilingual text to 86.4Mb. Second, we produced a program that can accurately predict the cipher system used to produce an observed cipher [53].

6 Catalog of Tools and Resources

In addition to resources in, and tools for, the target languages (§2.1, §2.2, §2.3), this project also created a variety of tools and resources for use in morphologically-complex languages in general, as well as some tools for use in the source language (esp. English). While rich syntactic and semantic annotations on the target side (e.g. Kinyarwanda and Malagasy) may not always be available in quantity, rich annotations on the source side can potentially be used in techniques like tree-to-string translation and multi-task learning.

6.1 Tools for processing and generating morphologically-rich and low-resource languages

- fast umorph, software for unsupervised induction of morphological grammars in agglutinative languages https://github.com/vchahun/fast_umorph
- morphogen, a package for translation into morphologically rich languages using synthetic translation options [59] https://github.com/eschling/morphogen
- morpholm is the implementation of the morphology-aware Bayesian language model [10] https://github.com/ldmt-muri/morpholm
- The CMU cross-lingual metaphor detector - a toolkit for identifying instances of figurative language in English and any other language for which a bilingual dictionary is available [64] https://github.com/ytsvetko/metaphor

6.2 Tools for generating and utilizing rich linguistic annotations and representations in the higher-resource language

- JAMR, an open-source parser for the Abstract Meaning Representation [30] https://github.com/jflanigan/jamr
- TreeTransductionTools for cdec, an implementation of tree-to-string transduction based translation and rule learning in the cdec framework https://github.com/pauldb89/worm
- English adjective supersense data and English adjective supersense classifier https://github.com/ytsvetko/adjective_supersense_classifier
- Highly re-entrant semantic graph corpus: http://amr.isi.edu/download/boygirl.tgz
- Bolinas, a package for graph processing based on Synchronous Hyperedge Replacement Grammars [12] http://www.isi.edu/publications/licensed-sw/bolinas/index.html
- GFL Tools for verifying, parsing, and evaluating annotations specified in the graph fragment language (Schneider et al., 2013) https://github.com/brendano/gfl_syntax/

- English to AMR: a program that parses English sentences into Abstract Meaning Representation `http://www.isi.edu/natural-language/software/`
- AMR to English: a program that generates English sentences from Abstract Meaning Representation `http://www.isi.edu/natural-language/software/`
- Multi-lingual Dependency Parsing: The source code is available at `https://github.com/yuanzh/TensorTransfer`
- Sparse encoding vectors `https://github.com/mfaruqui/sparse-coding`

References

[1] Waleed Ammar, Victor Chahuneau, Michael Denkowski, Greg Hanneman, Kenton Murray, Wang Ling, Austin Matthews Kenton, Yulia Tsvetkov, Nicola Segall, Chris Dyer, and Alon Lavie. The CMU machine translation systems at wmt 2013: Syntax, synthetic translation options, and pseudo-references. In *8th Workshop on Statistical Machine Translation*, page 70, 2013.

[2] Waleed Ammar, Chris Dyer, and Noah A Smith. Transliteration by sequence labeling with lattice encodings and reranking. In *Proceedings of the 4th Named Entity Workshop*, pages 66–70. Association for Computational Linguistics, 2012.

[3] Waleed Ammar, Chris Dyer, and Noah A Smith. Conditional random field autoencoders for unsupervised structured prediction. In *Advances in Neural Information Processing Systems*, pages 3311–3319, 2014.

[4] Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. Many languages, one parser. *arXiv preprint arXiv:1602.01595*, 2016.

[5] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract Meaning Representation for sembanking. In *Proc. ACL Linguistic Annotation Workshop*, 2013.

[6] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *In Proceedings of the 7th*

Linguistic Annotation Workshop and Interoperability with Discourse. Citeseer, 2013.

[7] F. Braune, D. Bauer, and K. Knight. Mapping between English strings and reentrant semantic graphs. In *Proc. LREC*, 2014.

[8] Shu Cai and Kevin Knight. Smatch: an evaluation metric for semantic feature structures. In *ACL (2)*, pages 748–752, 2013.

[9] Victor Chahuneau, Eva Schlinger, Noah A Smith, and Chris Dyer. Translating into morphologically rich languages with synthetic phrases. Association for Computational Linguistics, 2013.

[10] Victor Chahuneau, Noah A Smith, and Chris Dyer. Knowledge-rich morphological priors for bayesian language models. Association for Computational Linguistics, 2013.

[11] Desai Chen, Chris Dyer, Shay B Cohen, and Noah A Smith. Unsupervised bilingual pos tagging with markov random fields. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, pages 64–71. Association for Computational Linguistics, 2011.

[12] David Chiang, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, Bevan Jones, and Kevin Knight. Parsing graphs with hyperedge replacement grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 924–932, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[13] Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 176–181. Association for Computational Linguistics, 2011.

[14] Shay B Cohen, Dipanjan Das, and Noah A Smith. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 50–61. Association for Computational Linguistics, 2011.

[15] Chris Collins and Richard Kayne. *Syntactic Structures of the World's Languages*. New York University, New York, 2011.

[16] Q. Dou and K. Knight. Large scale decipherment for out-of-domain machine translation. In *Proc. EMNLP*, 2012.

[17] Q. Dou and K. Knight. Dependency-based decipherment for resource-limited machine translation. In *Proc. EMNLP*, 2013.

[18] Q. Dou, A. Vaswani, K. Knight, and C. Dyer. Unifying Bayesian inference and vector space models for improved decipherment. In *Proc. ACL*, 2015.

[19] Qing Dou, Ashish Vaswani, and Kevin Knight. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proc. EMNLP*, 2014.

[20] F. Drewes, H.-J. Kreowski, and A. Habel. Hyperedge replacement graph grammars. *Handbook of Graph Grammars*, 1:95–162, 1997.

[21] Matthew S. Dryer and Martin Haspelmath. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, 2013.

[22] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*, 2015.

[23] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of IBM model 2. Association for Computational Linguistics, 2013.

[24] Chris Dyer, Jonathan Clark, Alon Lavie, and Noah A Smith. Unsupervised word alignment with arbitrary features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 409–419. Association for Computational Linguistics, 2011.

[25] Chris Dyer, Kevin Gimpel, Jonathan H Clark, and Noah A Smith. The CMU-ARK German-English translation system. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 337–343. Association for Computational Linguistics, 2011.

[26] Chris Dyer, Noah A Smith, Graham Morehead, Phil Blunsom, and Abby Levenberg. The cmu-oxford translation system for the nist open machine translation 2012 evaluation. 2012.

[27] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004*, 2015.

[28] Jeffrey Flanigan, Chris Dyer, and Jaime G Carbonell. Large-scale discriminative training for statistical machine translation using held-out line search. In *HLT-NAACL*, pages 248–258, 2013.

[29] Jeffrey Flanigan, Chris Dyer, Noah A Smith, and Jaime G Carbonell. Generation from abstract meaning representation using tree transducers. 2016.

[30] Jeffrey Flanigan, Sam Thomson, Jaime G Carbonell, Chris Dyer, and Noah A Smith. A discriminative graph-based parser for the abstract meaning representation. 2014.

[31] Dan Garrette and Jason Baldridge. Type-supervised hidden markov models for part-of-speech tagging with incomplete tag dictionaries. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 821–831. Association for Computational Linguistics, 2012.

[32] Dan Garrette and Jason Baldridge. Learning a part-of-speech tagger from two hours of annotation. In *HLT-NAACL*, pages 138–147, 2013.

[33] Dan Garrette, Chris Dyer, Jason Baldridge, and Noah A Smith. Weakly-supervised bayesian learning of a ccg supertagger. Association for Computational Linguistics, 2014.

[34] Dan Garrette, Jason Mielens, and Jason Baldridge. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *ACL (1)*, pages 583–592, 2013.

[35] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[36] Kevin Gimpel and Noah A Smith. Generative models of monolingual and bilingual gappy patterns. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 512–522. Association for Computational Linguistics, 2011.

[37] Kevin Gimpel and Noah A Smith. Quasi-synchronous phrase dependency grammars for machine translation. In *Proceedings of the Conference on Em-*

pirical Methods in Natural Language Processing, pages 474–485. Association for Computational Linguistics, 2011.

[38] Kevin Gimpel and Noah A Smith. Concavity and initialization for unsupervised dependency parsing. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 577–581. Association for Computational Linguistics, 2012.

[39] Kevin Gimpel and Noah A Smith. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 221–231. Association for Computational Linguistics, 2012.

[40] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog 2.6.* Max Planck Institute for the Science of Human History, Jena, 2015.

[41] B. Jones, J. Andreas, D. Bauer, K-M. Hermann, and K. Knight. Semantics-based machine translation with hyperedge replacement grammars. In *Proc. COLING*, 2012.

[42] Bevan Jones, Jacob Andreas, Daniel Bauer, Karl Moritz Hermann, and Kevin Knight. Semantics-based machine translation with hyperedge replacement grammars. In *COLING*, pages 1359–1376, 2012.

[43] T. Kamimura and G. Slutzki. Transductions of DAGs and trees. *Math. Syst. Theory*, 15(3):225–249, 1982.

[44] Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. Modeling syntactic context improves morphological segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 1–9. Association for Computational Linguistics, 2011.

[45] Tao Lei, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. Low-rank tensors for scoring dependency structures. Association for Computational Linguistics, 2014.

[46] M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the World, Eighteenth edition*. SIL International, Dallas, Texas, 2015.

[47] Chu-Cheng Lin, Waleed Ammar, Chris Dyer, and Lori Levin. Unsupervised pos induction with word embeddings. *arXiv preprint arXiv:1503.06760*, 2015.

[48] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

[49] Austin Matthews, Waleed Ammar, Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Alon Lavie, and Chris Dyer. The CMU machine translation systems at WMT 2014. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*. Citeseer, 2014.

[50] Steven Moran, Daniel McCloy, and Richard Wright. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2014.

[51] Tahira Naseem and Regina Barzilay. Using semantic cues to learn syntax. In *AAAI*, 2011.

[52] Tahira Naseem, Regina Barzilay, and Amir Globerson. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics, 2012.

[53] Malte Nuhn and Kevin Knight. Cipher type detection. In *Proc. EMNLP*, 2014.

[54] K. Papineni, S. Roukos, T. Ward, and W. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[55] Elias Ponvert, Jason Baldridge, and Katrin Erk. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1077–1086. Association for Computational Linguistics, 2011.

[56] D. Quernheim and K. Knight. DAGGER: A toolkit for automata on directed acyclic graphs. In *Proc. FSMNLP*, 2012.

[57] D. Quernheim and K. Knight. Towards probabilistic acceptors and transducers for feature structures. In *Proc. ACL SSST Workshop*, 2012.

[58] S. Ravi and K. Knight. Deciphering foreign language. In *Proc. ACL*, 2011.

[59] Eva Schlinger, Victor Chahuneau, and Chris Dyer. morphogen: Translation into morphologically rich languages with synthetic phrases. *The Prague Bulletin of Mathematical Linguistics*, 100:51–62, 2013.

[60] Nathan Schneider, Brendan O'Connor, Naomi Saphra, David Bamman, Manaal Faruqui, Noah A Smith, Chris Dyer, and Jason Baldridge. A framework for (under) specifying dependency syntax without overloading annotators. *arXiv preprint arXiv:1306.2091*, 2013.

[61] Patrick Simianer, Stefan Riezler, and Chris Dyer. Joint feature selection in distributed stochastic learning for large-scale discriminative training in SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 11–21. Association for Computational Linguistics, 2012.

[62] Noah A. Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 354–362, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[63] Liang Sun, Jason Mielens, and Jason Baldridge. Parsing low-resource languages using Gibbs sampling for PCFGs with latent annotations. In *EMNLP*, pages 290–300, 2014.

[64] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258. Association for Computational Linguistics, 2014.

[65] Yulia Tsvetkov and Chris Dyer. Lexicon stratification for translating out-of-vocabulary words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 125–131, Beijing, China, July 2015. Association for Computational Linguistics.

[66] Yulia Tsvetkov, Chris Dyer, Lori Levin, and Archna Bhatia. Generating English determiners in phrase-based translation with synthetic translation options. In *8th Workshop on Statistical Machine Translation*, page 271, 2013.

[67] Yulia Tsvetkov, Florian Metze, and Chris Dyer. Augmenting translation models with simulated acoustic confusions for improved spoken language translation. Association for Computational Linguistics, 2014.

[68] Yuan Zhang, Regina Barzilay, and Amir Globerson. Transfer learning for constituency-based grammars. Association for Computational Linguistics, 2013.

[69] Yuan Zhang, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Greed is good if randomized: New inference for dependency parsing. 2014.

[70] Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. Learning to map into a universal POS tagset. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1368–1378. Association for Computational Linguistics, 2012.

[71] B. Zoph, M. Ghazvininejad, and K. Knight. How much information does a human translator add to the original? In *Proc. EMNLP*, 2015.